A BIT OF INFORMATION THEORY

LINDA PREISS ROTHSCHILD

Abstract. The aim of this article is to introduce the elements of the mathematics of information, pioneered by Claude Shannon at Bell Laboratories in 1948, to a general mathematical audience.

1. introduction

Information theory is an elegant mathematical construction dealing with the transmissions of symbols. The foundations and basic results of this theory were established by Claude Shannon,a mathematician at Bell Laboratories, and expounded in his groundbreaking paper, "A Mathematical Theory of Communication" [S1948] in 1948. The purpose of this article is to introduce mathematicians from many disciplines to some of the basic definitions and results of information theory in a precise language.

Here *information* is understood to be a sequence of symbols that are to be transmitted through some medium, called a *channel*. The fundamental concept of information theory is *entropy*, which represents the amount of uncertainty in a string of symbols, given some knowledge of the distribution of the symbols. As an example, suppose that one of the digits in $\{0, 1, 2, 3\}$ is to be communicated as a sequence of 0's and 1's, called *bits*. By representing these digits in binary,

$$0 \leftrightarrow 0, \quad 1 \leftrightarrow 1, \quad 2 \leftrightarrow 10, \quad 3 \leftrightarrow 11,$$

it is easy to see that if all digits are equally likely, then on average it takes 1.5 bits to communicate one digit in $\{0, 1, 2, 3\}$. Suppose, however, that the digits 0 and 1 are equally likely to occur, but that the digits 2 and 3 are two and four times as likely, respectively, as 0 or 1. Then the digits 0,1 each occur with probability 1/8, the digit 2 with probability 1/4, and the digit 3 with probability 1/2. By using the new representation

$$2 \leftrightarrow 0, \quad 3 \leftrightarrow 1, \quad 0 \leftrightarrow 10, \quad 1 \leftrightarrow 11$$

it will take on average $(1/2)(1)+(1/4)(1)+(1/8)(2)+(1/8)(2) = 1.25$ bits to communicate one of the digits.

Roughly speaking, the entropy of a packet of information is the average number of bits needed to represent it. In the second example above, more information about the distribution of the digits made it possible to use an average of only 1.25 bits instead of 1.5.

Date: 2015-1-13.

The roots of information theory are founded in discrete probability. In this exposition, I will follow an approach pioneered by R. W. Yeung [Y1991], [Y2008], in which some identities involving the entropy of finitely many random variables are translated into measure-theoretic language. The content here is not really original, although some results are stated more precisely and some proofs are given in more detail. The construction of subsets of a measure space associated to a vector of random variables is new.

2. Definition of entropy

Let's denote by P a probability measure living on some space $(\Omega, \mathcal{F})$, where Ω is the set of outcomes and $\mathcal{F}$ is a set of events, i.e. the collection of subsets of Ω on which P is defined. (In this article, $(\Omega, \mathcal{F})$, will remain in the background.) By a *discrete random variable*, I will mean a function

$$X : \Omega \to \mathcal{X}, \quad \mathcal{X} \subset \mathbb{R},$$

where $\mathcal{X}$ is discrete (countable or finite), and $X^{-1}(x) \in \mathcal{F}$ for all $x \in \mathcal{X}$. In this exposition, all random variables will be assumed to be discrete. The random variable X induces a probability measure p_X on $(\mathcal{X}, 2^{\mathcal{X}})$, where $2^{\mathcal{X}}$ is the set of all subsets of $\mathcal{X}$. The for $E \subset \mathcal{X}$, $p_X(E)$ is defined by

$$p_X(E) := \sum_{x \in E} P(\{\omega \in \Omega : X(\omega) = x\}).$$

For simplicity of notation, I shall omit the subscript X and write p for p_X where there is no risk of confusion. Also, I will follow the usual convention of writing $\{\omega \in \Omega : X(\omega) = x\}$ as $\{X = x\}$. If f is a real-valued function on $\mathcal{X}$, the *expected value* $E_p(f)$ is defined by

$$E_p(f) = \sum_{x \in \mathcal{X}} p(x) f(x).$$

Definition 2.1. *For a random variable X with values in the discrete subset $\mathcal{X} \subset \mathbb{R}$, its* entropy*, $H(X)$, is defined by*

$$H(X) := -\sum_{x \in \mathcal{X} : p(x) \neq 0} p(x) \log_2 p(x) \tag{2.1}$$

Since $x \mapsto \log_2 p(x)$ defines a random variable on $\mathcal{X}$ (omitting the points of probability 0), so the entropy of X is the expected value

$$H(X) = E_p(-\log_2 p).$$

Intuitively, the larger the entropy, the more uncertainty about the values of X. For example, suppose $\mathcal{X} = \Omega = \{1, 2, \ldots, N\}$ and P is the uniform distribution on Ω, i.e. $p(\omega) = 1/N$ for all $\omega \in \Omega$, then any bijection $X : \Omega \to \mathcal{X}$ satisfies $H(X) = \log_2 N$, since $p_X(x) = 1/N$ for all $x \in \mathcal{X}$. It is not hard to check, e.g. by using the method of Lagrange multipliers, that the minimum value of the function $p = (p_1, \ldots, p_N) \mapsto \sum_{\{p : p_j \neq 0\}} p_j \log_2 p_j$ subject to the constraint $\sum p_j = 1$ is achieved when $p_j = 1/N$ for all j. Hence, for the

given $\Omega, \mathcal{X}$, and p the maximum possible value of the entropy is achieved by any random variable for which $p(x)$ is constant, i.e. X is equally likely to take any values of $\mathcal{X}$.

The above argument also shows that for any probability space (Ω, P) and any random variable X with values in a finite set $\mathcal{X}$ consisting of N elements,

$$0 \leq H(X) \leq \log_2 N$$

The significance of $\log_2 N$ is that it is the number of digits needed to represent all integers between 0 and N in binary. If X is constant, $p(x) \in \{0, 1\}$ for all $x \in \mathcal{X}$; hence $H(X) = 0$. In this case there is no uncertainty about the possible value of X.

Remark 2.2. The justification for omitting those $x \in \mathcal{X}$ for which $p(x) = 0$ in the summand on the right in (2.1) is that $\lim_{y \to 0^+} y \log_2 y = 0$.

Remark 2.3. It is possible to have $H(X) = \infty$. To see this, take p_n satisfying $\sum_n p_n = 1$, but $\sum_n p_n \log_2 p_n = -\infty$. If X satisfies $P(\{X = n\}) = p_n$, then $H(X) = \infty$. Such a sequence p_n may be constructed, for example, by normalizing the sequence $a_n := 1/n(\log_2 n)^2$, $n = 2, 3, \ldots$, to sum to 1. For most of this article, we shall assume all entropies to be finite.

Remark 2.4. Entropy can also be defined in terms of a probability p on $\mathbb{N}$, the set of positive integers, by

$$H(p) := -\sum_{j \in \mathbb{N} : p(j) \neq 0} p(j) \log_2 p(j)$$

as in Shannon's original paper [S1948]. To see that this is a special case of (2.1), one can use a standard construction of a random variable $X : \Omega \to \mathcal{X}$ for which $p_X = p$ as follows. Take $\Omega := [0, 1] \subset \mathbb{R}$ with probability given by Lebesgue measure, and $\mathcal{X} := \mathbb{N}$. Let I_1be the subinterval of $[0, 1]$ given by $I_1 := [0, p(1)]$ and define $I_j := (p(j-1), p(j-1) + p(j)]$ for $j > 1$. Then $X : [0, 1] \to \mathbb{N}$ given by

$$X(\omega) := j \text{ for } \omega \in I_j$$

satisfies $P\{(X = j\} = p(j),\ j \in \mathbb{N}$.

3. Joint and conditional entropy, mutual information

If Y is another random variable taking values in a discrete set $\mathcal{Y}$, we let $p_Y(y) := P(\{Y = y\})$ and define the induced product probability, $p_{(X,Y)}$, on $\mathcal{X} \times \mathcal{Y}$ by

$$p_{(X,Y)}(x, y) = P(\{X = x \ \&\ Y = y\}),\ (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

As before, I shall drop the subscripts of p when the meaning is clear.

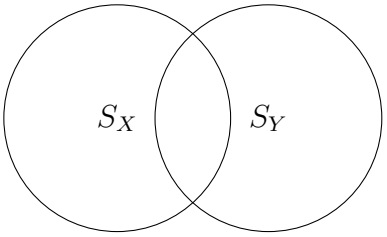
The *joint entropy* of X and Y may then be defined by

$$H(X, Y) := -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y} : p(x,y) \neq 0} p(x, y) \log_2 p(x, y),$$

or by $H(X,Y) = E_p(-\log_2 p)$, where p denotes the induced probability measure on $\mathcal{X}\times\mathcal{Y}$. Recall that the random variables X and Y are said to be *independent*, written $X \perp Y$ if $p(x,y) = p(x)p(y)$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Using the additivity of the $\log_2$ function, one can check

$$H(X,Y) = H(X) + H(Y) \iff X \perp Y.$$

This observation suggests exploring relationships between the joint entropy $H(X,Y)$ and the entropies of X and Y. Following the ideas in Yeung [Y2008], we will construct a finite set Ω', with subsets S_X, S_Y corresponding to X,Y respectively and a measure m defined on $\mathcal{F}' := 2^{\Omega'}$, the set of all subsets of Ω'. (Here, a *measure* means a nonnegative function m defined on $\mathcal{F}'$ and satisfying finite additivity, i.e. $m(S_1) + m(S_2) = m(S_1 \cup S_2)$ for any pair of disjoint sets in $\mathcal{F}'$.) To avoid undefined terms, I will assume that $H(X)$ and $H(Y)$ are both finite.



Let $\Omega' := S_X \cup S_Y$ and let $S_{(X,Y)}$ also denote this union. For any $S \subset \Omega'$ let S^c denote the complement of S in Ω'. We will choose S_X and S_Y in some way so that $S_X \cap S_Y^c$, $S_X \cap S_Y$, and $S_X^c \cap S_Y$ each contain just a single point. For this, we will take

$$S_X = \{1,2\},\ S_Y = \{2,3\},\ \text{and } \Omega' = S_{(X,Y)} = S_X \cup S_Y = \{1,2,3\}, \tag{3.1}$$

so that

$$S_X \cap S_Y^c = \{1\},\ S_X \cap S_Y = \{2\}, S_X^c \cap S_Y = \{3\}.$$

We want to assign the measure m on $(\Omega', \mathcal{F}')$ in such a way that

$$m(S_X) = H(X) \text{ and } m(S_Y) = H(Y),\ \text{and } m(S_{(X,Y)}) = H(X,Y). \tag{3.2}$$

I claim that there is a unique measure m defined on all subsets of Ω' satisfying (3.2). To check the claim, note first that by finite additivity, for any $S \subset \Omega'$,

$$m(S^c) = m(\Omega') - m(S).$$

Hence

$$m(S_X \cap S_Y^c) = m(S_Y^c) = m(\Omega') - m(S_Y) = H(X,Y) - H(Y) \tag{3.3}$$

and, similarly,

$$m(S_X^c \cap S_Y) = H(X,Y) - H(X)$$

From

$$m(S_{(X,Y)}) = m(S_X \cup S_Y) = m(S_X) + m(S_Y) - m(S_X \cap S_Y),$$

we have

$$m(S_X \cap S_Y) = m(S_X) + m(S_Y) - m(S_{(X,Y)}) = H(X) + H(Y) - H(X,Y). \tag{3.4}$$

Hence $m(\{1\}), m(\{2\})$ and $m(\{3\})$ are completely determined by assigning $m(S_X), m(S_Y)$, and $m(S_{(X,Y)})$ according to (3.2) . By finite additivity, m is uniquely determined, which proves the claim.

Remark 3.1. The choice of sets given by (3.1) is arbitrary. Any sets S_X, S_Y for which $S_X \cap S_Y, S_X^c \cap S_Y$ and $S_X \cap S_Y^c$ are all distinct would also give rise to a unique measure (on an algebra of subsets of $S_X \cup S_Y$) satisfying (3.2).

To interpret the measures of the intersections , consider first $H(X,Y) - H(Y)$. We understand this as the joint uncertainty of the random variables X and Y minus the uncertainty of Y, i.e. given that Y is known. This suggests conditional probability. Recall that the *conditional probability* of $x \in \mathcal{X}$ given $y \in \mathcal{Y}$ is defined by $p(x|y) := p(x,y)/p(y)$ if $p(y) \neq 0$. (By definition, $p(y) = 0$ implies $p(x,y) = 0$.) If we define the *conditional entropy* $H(X|Y)$ of X and Y as the negative of the expected value (with respect to $p_{(X,Y)}$) of the function $\mathcal{X} \times \mathcal{Y} \ni (x,y) \mapsto \log_2(p(x|y)$, then

$$H(X|Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}:p(x,y)\neq 0} p(x,y)\log_2 p(x|y) \tag{3.5}$$

The equality $\sum_{(x,y):p(x,y)\neq 0} p(x,y) = p(y)$ holds for any $y \in \mathcal{Y}$. By using this equality and the additive properties of $\log_2$, it is easy to check that

$$H(X|Y) = H(X,Y) - H(Y) \tag{3.6}$$

and hence

$$H(X|Y) = m(S_X \cap S_Y^c).$$

Finally, consider

$$m(S_X \cap S_Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

where the last 2 equalities follow from (3.6). This quantity is the uncertainty of X minus the uncertainty of X given Y (or vice versa). To understand this better, recall that $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are *independent* if $p(x,y) = p(x)p(y)$, and X and Y are *independent* as random variables if x and y are independent for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. From (3.5) it is clear that X and Y are independent if and only if $H(X) = H(X|Y)$, i.e. if no information is gained about X from knowing Y. This motivates the following definition.

Definition 3.2. *If X and Y are random variables as above with $H(X)$, $H(Y)$ their respective entropies and $H(X,Y)$ their joint entropy, their* mutual information $I(X;Y)$ *is defined by*

$$I(X;Y) := H(X) + H(Y) - H(X,Y).$$

This definition gives an interpretation of the remaining underlying set:

$$I(X;Y) = m(S_X \cap S_Y).$$

We end with a few more observations concerning mutual information. It is easy to check that

$$I(X;Y) = \sum_{((x,y)\in\mathcal{X}\times\mathcal{Y}:p(x,y)\neq 0} p(x,y)\log_2(\frac{p(x,y)}{p(x)p(y)}), \tag{3.7}$$

and $I(X;Y)$ is the expected value of the function $(x,y) \mapsto \log_2(p(x,y)/p(x)p(y))$. It is less immediate that $I(X;Y) \geq 0$ (from which the positivity of the measure m will follow). Intuitively, $H(X|Y)$ should not be greater than $H(X)$, since the uncertainty of the values of X cannot be increased with knowledge of another random variable Y. Here is a formal statement and proof of this fact.

Proposition 3.3. *Mutual information is nonnegative, i.e.* $I(X;Y) \geq 0$. *Equivalently,* $H(X|Y) \leq H(X)$. *Hence conditioning one random variable on another can only decrease entropy. Equality holds if and only if the random variables are independent.*

Proof. Since $\log_2 z \equiv \frac{1}{\ln 2}\ln z$, , where ln denotes the natural log, it suffices to prove the proposition with $\log_2$ replaced by ln in (3.7). Let's make use of the well-known inequality

$$\ln z \geq 1 - \frac{1}{z} \text{ for } z > 0$$

with equality if and only if $z = 1$. (This inequality may be proved by showing that the function $z \mapsto \ln z - (1 - \frac{1}{z})$ reaches a global minimum of 0 at $z = 1$.) Substituting the inequality into (3.7) (with $\log_2$ replaced by ln), we have

$$\begin{aligned}\sum_{(x,y):p(x,y)\neq 0} p(x,y)\ln\left(\frac{p(x,y)}{p(x)p(y)}\right) &\geq \sum_{(x,y):p(x,y)\neq 0} p(x,y)\left(1 - \frac{p(x)p(y)}{p(x,y)}\right) \\ &= \sum_{x,y:p(x,y)\neq 0} p(x,y) - \sum_{(x,y):p(x,y)\neq 0} p(x)p(y) \geq 0,\end{aligned} \tag{3.8}$$

where the last inequality follows from $\sum_{x,y:p(x,y)\neq 0} p(x,y) = 1$, while

$$\sum_{x,y:p(x,y)\neq 0} p(x)p(y) \leq \sum_{x,y} p(x)p(y) = 1$$

Hence $I(X;Y) \geq 0$. For the last statement, note that equality holds in the first line if and only if $\frac{p(x,y)}{p(x)p(y)} = 1$ when $p(x,y) \neq 0$. □

Remark 3.4. The sum $\sum_{x,y:p(x,y)\neq 0} p(x)p(y)$ may be strictly less than 1 (although this possibility is not always addressed in the literature). As an example, suppose

$$p(1,0) = p(1,1) = 1/4,\ p(0,1) = 0,\ p(0,0) = 1/2,$$

so that $p(\{x=1\}) = 1/2,\ p(\{y=1\}) = 1/4$. Then

$$\sum_{x,y:p(x,y)\neq 0} p(x)p(y) = p(\{x=1\})p(\{y=0\})+ \tag{3.9}$$
$$p(\{x=1\})p(\{y=1\}) + p(\{x=0\})p(\{y=0\}) = (1/2)(3/4) + 1/8 + 3/8 = 7/8.$$

Remark 3.5. The non negativity of $I(X;Y)$ will also follow from a more general result proved in Section 6 below.

4. Entropy among n random variables

A measure space containing three points does not seem very exciting, but the theory can be expanded to a set of n random variables $X_1, \ldots, X_n$ with values in discrete sets $\mathcal{X}_1, \ldots, \mathcal{X}_n$. To fix notation, let

$$\mathcal{X} := \mathcal{X}_1 \times \ldots \times \mathcal{X}_n,$$

and for $F = \{i_1, \ldots, i_k\} \subset \{1, 2, \ldots, n\}$, write

$$X_F := (X_{i_1}, \ldots, X_{i_k}) \text{ and } \mathcal{X}_F = \mathcal{X}_{i_1} \times \ldots \mathcal{X}_{i_k}, \tag{4.1}$$

For $\mathbf{x}_F = (\mathbf{x}_{i_1}, \ldots \mathbf{x}_{i_k}) \in \mathcal{X}_F$, let

$$p(\mathbf{x}_F) := P(\{X_{I_1} = \mathbf{x}_{i_1}\}, \ldots, \{X_{I_k} = \mathbf{x}_{i_k}\}) \tag{4.2}$$

As before, p defines a probability measure $p_{\mathcal{X}_F}$ on $\mathcal{X}_F$, which will again be denoted by p if there is no ambiguity. The *joint entropy*, $H(X_F)$, is defined by

$$H(X_F) := \sum_{\mathbf{x}_F \in \mathcal{X}_F : p(\mathbf{x}_F)\neq 0} p(\mathbf{x}_F) \log_2(p(\mathbf{x}_F))$$

As in the case of two random variables, the joint entropy can be expressed in terms of expected value of the function $\mathcal{X}_F \ni \mathbf{x}_F \mapsto -\log_2(p(\mathbf{x}_F))$. For the rest of this section, we shall assume all entropies are finite.

If F_1 and F_2 are subsets of $\{1, 2, \ldots, n\}$, the *conditional entropy*, $H(X_{F_1}|X_{F_2})$, is defined by

$$H(X_{F_1}|X_{F_2}) := -\sum p(\mathbf{x}_{F_1\cup F_2}) \log_2 \big(\frac{p(\mathbf{x}_{F_1\cup F_2})}{p(\mathbf{x}_{F_2})}\big),$$

where the sum is taken over $\{\mathbf{x}_{F_1\cup F_2} \in \mathcal{X}_{F_1\cup F_2} : p(\mathbf{x}_{F_1\cup F_2}) \neq 0\}$. (By definition, if $p(\mathbf{x}_{F_1}) = 0$ or $p(\mathbf{x}_{F_2}) = 0$, then $p(\mathbf{x}_{F_1\cup F_2}) = 0$, so that the right hand side of the equation is always defined.) The fraction $\big(\frac{p(\mathbf{x}_{F_1\cup F_2})}{p(\mathbf{x}_{F_2})}\big)$ is the conditional probability of $\mathbf{x}_{F_1\cup F_2}$ given $\mathbf{x}_{F_2}$, written $p(\mathbf{x}_{F_1\cup F_2}|\mathbf{x}_{F_2})$. Then $H(X_{F_1}|X_{F_2})$ may be expressed as the expected value of the function $\mathbf{x}_{F_1\cup F_2} \mapsto -\log_2(p(\mathbf{x}_{F_1\cup F_2}|\mathbf{x}_{F_2}))$. The identity

$$H(X_{F_1}|X_{F_2}) = H(X_{F_1\cup F_2}) - H(X_{F_2}), \tag{4.3}$$

follows from the definition, as in the case of 2 random variables.

The *mutual information* $I(X_{F_1}; X_{F_2})$ may be defined as for the case of 2 random variables by

$$I(X_{F_1}; X_{F_2}) := \sum p(\mathbf{x}_{F_1 \cup F_2}) \log_2 \big(\frac{p(\mathbf{x}_{F_1 \cup F_2})}{p(\mathbf{x}_{F_1}) p(\mathbf{x}_{F_2})}\big),$$

with the sum taken over $\{\mathbf{x}_{F_1 \cup F_2} \in \mathcal{X}_{F_1 \cup F_2} : p(\mathbf{x}_{F_1 \cup F_2}) \neq 0\}$ as before. However, if $F_3 \subset \{1, 2, \ldots, n\}$ is another set of indices, it is natural to define the *conditional mutual information*, $I(X_{F_1}; X_{F_2} | X_{F_3})$ by

$$I(X_{F_1}; X_{F_2} | X_{F_3}) := \sum p(\mathbf{x}_{F_1 \cup F_2 \cup F_3}) \log_2 \big(\frac{p(\mathbf{x}_{F_1 \cup F_2} | \mathbf{x}_{F_3})}{p(\mathbf{x}_{F_1} | \mathbf{x}_{F_3}) p(\mathbf{x}_{F_2} | \mathbf{x}_{F_3})}\big), \tag{4.4}$$

with the sum taken over $\{\mathbf{x}_{F_1 \cup F_2 \cup F_3} \in \mathcal{X}_{F_1 \cup F_2 \cup F_3} : p(\mathbf{x}_{F_1 \cup F_2 \cup F_3}) \neq 0\}$.

We may use the additivity of the $\log_2$ function to check the identities

$$I(X_{F_1}; X_{F_2}) = H(X_{F_1}) + H(X_{F_2}) - H(X_{F_1 \cup F_2}) \quad \text{and} \tag{4.5}$$

$$I(X_{F_1}; X_{F_2} | X_{F_3}) = H(X_{F_1} | X_{F_3}) + H(X_{F_2} | X_{F_3}) - H(X_{F_1 \cup F_2} | X_{F_3}) \tag{4.6}$$

The notions of joint entropy and conditional entropy can be realized as special cases of conditional mutual information as follows. Taking $F_2 = F_1$ in (4.6),

$$I(X_{F_1}; X_{F_1} | X_{F_3}) = H(X_{F_1} | X_{F_3}),$$

since all three terms on the right become identical.

5. Measure-theoretic identities

Here $X_1, \ldots X_n$ will denote random variables as before. To avoid undefined terms, we shall again assume that $H(X_i) < \infty$, $i = 1, \ldots, n$. In order to prove an analog of the results in Section 3, we will construct a finite space Ω', and a signed measure m defined on $\mathcal{A} = 2^{\Omega'}$, the algebra of all subsets of Ω'. By a (finite) signed measure, I shall mean a function $m : \mathcal{A} \to \mathbb{R}$ satisfying finite additivity:

$$m(S_1 \cup S_2) = m(S_1) + m(S_2) - m(S_1 \cap S_2),$$

for any $S_1, S_2 \in \mathcal{A}$. A signed measure necessarily satisfies $m(\emptyset) = 0$, but $S_1 \subset S_2$ need not imply $m(S_1) \leq m(S_2)$.

Following the ideas used in the case $n = 2$, we shall associate to each X_j a set S_j and put $\Omega' = \bigcup_{j=1}^n S_i$. Then every nonempty subset of Ω' can be expressed as a union sets in $\mathcal{B}$, where

$$\mathcal{B} = \{S \subset \Omega' : S = \bigcap_{j=1}^n T_j, \quad \text{where} \;\; T_j \in \{S_j, S_j^c\}, \; T_j = S_j \text{ for at least one } j\}$$

There are $2^n - 1$ such possibly nonempty sets in $\mathcal{B}$, (since $\bigcap_{j=1}^n S_j^c = \emptyset$). If the S_j are chosen so that the $2^n - 1$ sets in $\mathcal{B}$ are disjoint and nonempty, then any map $m : \mathcal{B} \to \mathbb{R}$

extends uniquely to a signed measure m on $(\Omega', \mathcal{A})$. Under these conditions Ω' must contain at least $2^n - 1$ elements, which motivates our choice:

$$\Omega' = \{K \in \mathbb{Z} : 1 \leq K \leq 2^n - 1\} \text{ and } \mathcal{A} = 2^{\Omega'}. \tag{5.1}$$

In order to associate each random variable X_j, $1 \leq j \leq n$ with a subset $S_j \subset \Omega'$, it is convenient to use the binary representation $b(j)$ of j (which has $2^n - 1$ or fewer nonzero digits). For $1 \leq K \leq 2^n - 1$, let $b_K(j)$ denote the Kth digit of $b(j)$ (from the right) and define S_j by

$$S_j := \{K \in \Omega' : b_j(K) = 1\}.$$

For example,

$$S_1 = \{J,\ 1 \leq J \leq 2^n - 1 : b_1(J) = 1\},$$

which is the set of all odd numbers from 1 to $2^n - 1$. Each $S \in \mathcal{B}$ contains a unique integer J with $1 \leq J \leq 2^n - 1$, since all its binary digits are specified. It is clear that a signed measure m on $(\Omega', \mathcal{A})$ is uniquely determined by any assignment $J \to m(\{J\}) \in \mathbb{R}$.

For $F = \{i_1, \ldots, i_k\} \subset \{1, 2 \ldots n\}, F \neq \emptyset$, we recall the notation $X_F := (X_{i_1}, \ldots, X_{i_k})$ of Section 4, and define the corresponding set $S_F \in \mathcal{A}$ by

$$S_F = \bigcup_{j \in F} S_j = \{K \in \Omega' : b_j(K) = 1 \text{ for some } j \in F\},$$

and note that $S_F^c = \{K \in \Omega' : b_j(K) = 0 \text{ for all } j \in F\}$. For simplicity, we shall retain the notation S_i for $S_{\{i\}}$.

Let $\mathcal{U} = \{S_F \subset \Omega' : F \subset \{1, 2 \ldots n\}, F \neq \emptyset\}$, and observe that $\mathcal{U}$ is closed under all unions, since

$$S_{F_1} \cup S_{F_2} = S_{F_1 \cup S_{F_2}} \text{ for (nonempty) } F_1, F_2 \subset \{1, 2, \ldots, n\} \tag{5.2}$$

Note that $\mathcal{U} \subset \mathcal{A}$ contains $2^n - 1$ elements, while the order of $\mathcal{A}$ is $2^{2^n - 1}$.

Theorem 5.1. *Let $X_1, \ldots X_n$ be random variables with values in a discrete set $\mathcal{X} = \mathcal{X}_1 \times \ldots \mathcal{X}_n$. Let $(\Omega', \mathcal{A})$ be the pair defined by (5.1) and $S_F \in \mathcal{A}$ as in (5.5). Then the mapping*

$$m : S_F \mapsto H(X_F),$$

defined for all nonempty $F \subset \{1, 2 \ldots n\}$, extends uniquely to a signed measure m on $(\Omega', \mathcal{A})$ satisfying the following for any $F_1, F_2, F_3 \subset \{1, 2 \ldots n\}$:

1. $H(X_{F_1}, X_{F_2}) = m(S_{F_1} \cup S_{F_2}) = m(S_{F_1 \cup F_2})$
2. $H(X_{F_1} | X_{F_2}) = m(S_{F_1} \cap S_{F_2}^c)$
3. $I(X_{F_1}; X_{F_2}) = m(S_{F_1} \cap S_{F_2})$
4. $I(X_{F_1}; X_{F_2} | X_{F_3}) = m(S_{F_1} \cap S_{F_2} \cap S_{F_3}^c)$

Theorem 5.1 gives a method of translating the notions of joint entropy, conditional entropy, and mutual information into the more usual (for mathematicians) set-theoretic and measure-theoretic notions of union, complements, intersection, and finite additivity

for signed measures. As a consequence, one can prove so-called chain rules. For instance, the identity

$$H(X_1,\dots,X_k)=\sum_{i=1}^{k} H(X_i|X_{i-1},X_{i-2},\dots,X_1) \tag{5.3}$$

is easily checked by noting that the right hand side becomes a collapsing sum:

(5.4)

$$\sum_{i=1}^{k} H(X_i|X_{i-1},X_{i-2},\dots,X_1)=\sum_{i=1}^{k}\Big(m(S_{\{i\}}\cup S_{\{i-1\}}\cup\ldots\cup S_{\{1\}})-m(S_{\{i-1\}}\cup\ldots\cup S_{\{1\}})\Big)=$$
$$m(S_{\{i\}}\cup S_{\{i-1\}}\cup\ldots\cup S_{\{1\}})=H(X_1,\dots,X_k).$$

The proof of Theorem 5.1 will use the following construction for the signed measure m.

Lemma 5.2. *Let Ω', $\mathcal{U}$ be as above, and for $F\subset\{1,\dots,n\}$, let $m:\mathcal{U}\to[0,\infty)$ be defined by*

$$m(S_F)=H(X_F)\in[0,\infty) \text{ for } S_F\in\mathcal{U} \tag{5.5}$$

Then for all $J\in\Omega'$ there is a unique real number $m(\{J\})$ such that for any $F\subset\{1,2,\dots n\}$ nonempty,

$$m(S_F)=\sum_{J\in S_F} m(\{J\}). \tag{5.6}$$

Hence $m|\mathcal{U}$ extends to a unique signed measure m on $(\Omega',2^{\Omega'})$.

Proof. Let $\mathcal{D}\subset\Omega'$ be the set of integers J for which $m(\{J\})$ is uniquely determined by (5.5). The lemma will follow by proving $\mathcal{D}=\Omega'$. For $F\subset\{1,\dots,n\}$, $F\neq\emptyset$, let $J_F\in\Omega'$ be uniquely determined by

$$b_j(J_F)=0 \iff j\in F,$$

and note that for every $J\in\Omega'$ there is a unique proper subset $F\subset\{1,\dots,n\}$ such that $J=J_F$. By definition, J_F is in S_F^c and its binary representation is the unique element in S_F^c with exactly $|F|$ zeros. For $0\le\ell\le n$, let

$$\Omega'_\ell=\{J_F:|F|\ge\ell\},$$

so that

$$\emptyset=\Omega'_n\subset\Omega'_{n-1}\subset\ldots\subset\Omega'_0=\Omega'.$$

For any $J_F\in\Omega'_{n-1}$, we observe that $\{J_F\}=S_F^c$, so that $m(\{J_F\})=m(\Omega')-m(S_F)$. Hence $\Omega'_{n-1}\subset\mathcal{D}$. Now assume by induction that

$$\Omega'_{n-j}\subset\mathcal{D} \text{ for some } j,\ 1\le j\le n-1,$$

and let $J = J_F$ with $|F| = n - (j+1)$. To complete the the inductive proof, it suffices to show $J \in \mathcal{D}$. Since J is the unique element in S_F^c whose binary representation has exactly $n-(j+1)$ zeroes, it follows that

$$S_F^c \setminus \{J\} \subset \Omega'_{n-j},$$

so that $J \in \mathcal{D}$, which completes the proof of the lemma. □

We now prove Theorem 5.1.

Proof. Using Lemma 5.2, for any $S \subset \Omega'$ define $m(S)$ by

$$m(S) := \sum_{x \in S} m(\{x\}), \tag{5.7}$$

which is consistent with the definition of $m(S_F)$ given by (5.6). To complete the proof, we need to check (1) through (4). The equality (1) follows from the definition of S_F and (5.2). Since

$$m(S_{F_1} \cap S_{F_2}^c) = m(S_{F_1 \cup F_2}) - m(S_{F_2}),$$

the equality (2) follows from (4.3). To prove (4) we may use the additivity of m to check that

$$m(S_{F_1} \cap S_{F_2} \cap S_{F_3}^c) = m(S_{F_1} \cap S_{F_3}^c) + m(S_{F_1} \cap S_{F_3}^c) - m(S_{F_1 \cup F_2} \cap S_{F_3}^c),$$

which by (4.6), together with (1) and (2), equals $I(X_{F_1}; X_{F_2}|X_{F_3})$. The proof of (3) is the same, using (4.5). This completes the proof of Theorem 5.1. □

6. Positive and negative values of m

It turns out that all the quantities (1) through (4) of Theorem 5.1 are nonnegative as we will now check. It is clear that (1) is nonnegative, since it is a sum of nonnegative terms. Also, since (2) and (3) are special cases of (4), it suffices to prove the non negativity of (4). We'll begin with the log-sum inequality.

Proposition 6.1. *(Log-sum inequality) Let $a_1, \ldots, a_k$, $b_1, \ldots, b_k$ be nonnegative real numbers, satisfying $\sum_i a_i \neq 0$. Then*

$$\sum_{\{i: a_i \neq 0\}} a_i \log_2(\frac{a_i}{b_i}) \geq (\sum_i a_i) \log_2 \frac{\sum_i a_i}{\sum_i b_i}, \tag{6.1}$$

with equality if and only if there exists $r > 0$ such that $a_i = r b_i$, for all nonzero b_i. Here we have adopted the usual convention $\log_2(\frac{c}{0}) = \infty$ for $c > 0$.

Proof. By assumption, there is at least one nonzero a_i. If $b_i = 0$ for any i for which $a_i \neq 0$, the left hand side of (6.1) is infinity, and the inequality also holds. Hence we may

assume a_i and b_i are positive for the terms on the left hand side of the inequality (and also $\sum_{i:a_i\neq 0} b_i > 0$). Under this assumption, we may write the left hand side of (6.1) as

$$(\sum_j b_j)\Big(\sum_i (\frac{b_i}{\sum_j b_j})\frac{a_i}{b_i}\log_2(\frac{a_i}{b_i})\Big).$$

Since the function $x \mapsto x\log_2 x$ is strictly convex, for any positive real numbers $t_1, \dots t_k, x_1, \dots x_k$ satisfying $\sum_i t_i = 1$, we have

$$\sum_i t_i x_i \log_2 x_i \geq (\sum_i t_i x_i)\log_2(\sum_i t_i x_i), \tag{6.2}$$

with equality if and only if there exists $r > 0$ such that $x_i = r$ for all i. Take $t_i := \frac{b_i}{\sum_i b_i}$ and $x_i := \frac{a_i}{b_i}$ in (6.2) to obtain

$$\sum_{\{i:a_i\neq 0\}} a_i \log_2(\frac{a_i}{b_i}) \geq (\sum_i a_i)\log_2 \frac{\sum_i a_i}{\sum_{i:a_i\neq 0} b_i} \tag{6.3}$$

with equality holding if and only if a_i/b_i is constant in i for $a_i \neq 0$. This proves the proposition if there is no i for which $b_i \neq 0$ but $a_i = 0$. If such an i does exist, then

$$\sum_{\{i:a_i\neq 0\}} b_i < \sum_i b_i$$

so that strict inequality holds in (6.1), which completes the proof in this case also. □

Using the log-sum inequality, we may prove that (4) of Theorem 5.1 is nonnegative. The definition (4.4) may be rewritten as

$$I(X_{F_1}; X_{F_2}|X_{F_3}) = \sum p(\mathbf{x}_{F_1\cup F_2\cup F_3})\log_2\big(\frac{p(\mathbf{x}_{F_1\cup F_2\cup F_3})}{p(\mathbf{x}_{F_1}|\mathbf{x}_{F_3})p(\mathbf{x}_{F_2\cup F_3})}\big),$$

with the sum taken over $\{\mathbf{x}_{F_1\cup F_2\cup F_3} \in \mathcal{X}_{F_1\cup F_2\cup F_3} : p(\mathbf{x}_{F_1\cup F_2\cup F_3}) \neq 0\}$. Now apply the log-sum inequality with

$$\begin{aligned} \{a_i\} &\longleftrightarrow \{p(\mathbf{x}_{F_1\cup F_2\cup F_3}) \text{ for } \mathbf{x}_{F_1\cup F_2\cup F_3} \in \mathcal{X}_{F_1\cup F_2\cup F_3}\}, \\ \{b_i\} &\longleftrightarrow \{p(\mathbf{x}_{F_1}|\mathbf{x}_{F_3})p(\mathbf{x}_{F_2\cup F_3}) \text{ for } \mathbf{x}_{F_1\cup F_2\cup F_3} \in \mathcal{X}_{F_1\cup F_2\cup F_3}\}, \end{aligned} \tag{6.4}$$

where the indices i on the left side of (6.4) correspond to those $\mathbf{x}_{F_1\cup F_2\cup F_3} \in \mathcal{X}_{F_1\cup F_2\cup F_3}$ with $p(\mathbf{x}_{F_1\cup F_2\cup F_3}) \neq 0$ on the right side. Since

$$\sum_{\{\mathbf{x}_{F_1\cup F_2\cup F_3}: p(\mathbf{x}_{F_1\cup F_2\cup F_3})\neq 0\}} p(\mathbf{x}_{F_1\cup F_2\cup F_3}) = 1,$$

we have

$$I(X_{F_1}; X_{F_2}|X_{F_3}) \geq \log_2\Big(\frac{1}{\sum p(\mathbf{x}_{F_1}|\mathbf{x}_{F_3})p(\mathbf{x}_{F_2\cup F_3})}\Big),$$

where again it is important to note that the sum in the denominator on the right is taken only over those $\mathbf{x}_{F_1\cup F_2\cup F_3}$ satisfying $p(\mathbf{x}_{F_1\cup F_2\cup F_3}) \neq 0$. To prove $I(X_{F_1}; X_{F_2}|X_{F_3}) \geq 0$, it suffices to prove

$$\sum_{p(\mathbf{x}_{F_1\cup F_2\cup F_3})\neq 0} \frac{p(\mathbf{x}_{F_1\cup F_3})}{p(\mathbf{x}_{F_3})} p(\mathbf{x}_{F_2\cup F_3}) \leq 1 \tag{6.5}$$

To prove (6.5), let's first assume that the sets of indices $F_i, i = 1, 2, 3$ are pairwise disjoint. Under this assumption any $\mathbf{x}_{F_1\cup F_2\cup F_3} \in \mathcal{X}_{F_1\cup F_2\cup F_3}$ may written

$$\mathbf{x}_{F_1\cup F_2\cup F_3} = (\mathbf{x}_{F_1}, \mathbf{x}_{F_2}, \mathbf{x}_{F_3})$$

We estimate the left hand side of (6.5) by summing over all $(\mathbf{x}_{F_1}, \mathbf{x}_{F_2}, \mathbf{x}_{F_3})$ for which $p(\mathbf{x}_{F_3}) \neq 0$, which will give an upper estimate, since it includes all terms for which $p(\mathbf{x}_{F_1}, \mathbf{x}_{F_2}, \mathbf{x}_{F_3}) \neq 0$. By summing first over $\mathbf{x}_{F_1}$ and $\mathbf{x}_{F_2}$ with $\mathbf{x}_{F_3}$ fixed (and satisfying $p(\mathbf{x}_{F_3}) \neq 0$), we obtain the desired upper estimate:

$$\sum_{\{p(\mathbf{x}_{F_1\cup F_2\cup F_3})\neq 0\}} \frac{p(\mathbf{x}_{F_1\cup F_3})}{p(\mathbf{x}_{F_3})} p(\mathbf{x}_{F_2\cup F_3}) \leq \sum_{\{\mathbf{x}_{F_3}: p(\mathbf{x}_{F_3})\neq 0\}} \Big(\sum_{\mathbf{x}_{F_2}} p(\mathbf{x}_{F_2}, \mathbf{x}_{F_3})\Big) \sum_{\mathbf{x}_{F_1}} p(\mathbf{x}_{F_1}|\mathbf{x}_{F_3}) = 1. \tag{6.6}$$

This proves the non negativity of (4) in the case that the indices in the F_i are disjoint. For the general case, we replace the F_i by pairwise disjoint F_i' as follows.

$$F_3' := F_3,\ F_2' := F_2 \cap F_3^c,\ F_1' := F_1 \cap F_2^c \cap F_3^c.$$

Then

$$F_1' \cup F_2' \cup F_3' = F_1 \cup F_2 \cup F_3,\ F_2' \cup F_3' = F_2 \cup F_3,\ F_1' \cup F_3' \subset F_1 \cup F_3.$$

Hence

$$p(\mathbf{x}_{F_3}) = p(\mathbf{x}_{F_3'}),\ p(\mathbf{x}_{F_2\cup F_3}) = p(\mathbf{x}_{F_2'\cup F_3'}), \text{and } p(\mathbf{x}_{F_1\cup F_3}) \leq p(\mathbf{x}_{F_1'\cup F_3'}).$$

so that

$$\sum_{\{p(\mathbf{x}_{F_1\cup F_2\cup F_3})\neq 0\}} \frac{p(\mathbf{x}_{F_1\cup F_3})}{p(\mathbf{x}_{F_3})} p(\mathbf{x}_{F_2\cup F_3}) \leq \sum_{\{p(\mathbf{x}_{F_1'\cup F_2'\cup F_3'})\neq 0\}} \frac{p(\mathbf{x}_{F_1'\cup F_3'})}{p(\mathbf{x}_{F_3'})} p(\mathbf{x}_{F_2'\cup F_3'}) = 1$$

This proves the upper bound in the general case. The proof of the positivity of (4), and hence (2), and (3) also is complete.

7. Possible negativity of the measure m

Although the measures of all the relevant sets in Theorem 5.1 are nonnegative, the signed measure m may actually take negative values if $n \geq 3$. As with many examples, to find such an instance we choose three random variables that are pairwise independent but not independent as a triple.

Example 7.1. Let X_1 and X_2 be independent random variables with values in $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$, each taking 0 and 1 with probability .5. Let $X_3 = (X_1 + X_2) \mod 2$. (In the language of computer science, X_3 is the *XOR* function of X_1 and X_2.) Following the notation of Theorem 5.1,

$$F = \{1, 2, 3\}, S_{\{1\}} = \{1, 3, 5, 7\}, S_{\{2\}} = \{2, 3, 6, 7\}, \text{ and } S_{\{3\}} = \{4, 5, 6, 7\}.$$

It is easy to check that

$$H(X_1) = H(X_2) = H(X_3) = 1,$$

and by pairwise independence,

$$H(X_1, X_2) = H(X_1, X_3) = H(X_2, X_3) = 2.$$

By using the chain rule(5.3), we have

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) = 1 + 1 + 0 = 2,$$

since X_2 is independent of X_1, and X_3 is completely determined by X_1 and X_2. Also by pairwise independence

$$H(X_1|X_2) = H(X_2|X_3) = H(X_3|X_1) = 1,$$

and hence by Theorem 5.1 (2) and the additivity property of m,

$$m(S_{\{1\}} \cap S_{\{2\}}^c) + m(S_{\{2\}} \cap S_{\{3\}}^c) + m(S_{\{3\}} \cap S_{\{1\}}^c) = m(\{1, 5\}) + m(\{2, 3\}) + m(\{4, 6\}) = 3,$$

so that

$$m(S_{\{1\}} \cap S_{\{2\}} \cap S_{\{3\}}) = (m(\{7\}) = m(\{1, 2, 3, 4, 5, 6, 7\}) - m(\{1, 2, 3, 4, 5, 6\}) = 2 - 3 = -1.$$

Hence m takes negative values in this example, although all the quantities in Theorem 5.1 (1) through (4) are nonnegative.

8. Closing remarks

I have touched here only on some of the basic properties of entropy in the discrete setting. The textbooks [CT2006] and [Y2008] offer much further theory as well as many interesting applications to the transmission of data. On the theoretical side, for instance, it is useful to study entropy in the situation where the random variables are related via a Markov chain. For applications, the notion of mutual information is used to define the capacity of a channel, giving a theoretic upper limit for the rate at which information can be transmitted.

On a personal note, about ten years ago I decided to learn some of the mathematics involved in the communication and wireless revolution, so I went to a marvelous graduate course on information theory given by Alon Orlitsky, a professor of computer engineering at UCSD. Unfortunately, I was too busy that quarter to continue going to the course. A few years later, I tried again to start learning information theory together with one of my colleagues (whose mathematical research was as far from information theory as mine). We began at the first chapter (of a previous edition) of the textbook [CT2006], but were

deterred by some imprecision in the presentation and gave up the project. More recently I saw that R. W. Yeung, a professor of information engineering at the Chinese University of Hong Kong, was giving a course on information theory as a MOOC (Massive Open Online Course) on the Coursera platform. His approach to entropy identities via measure theory inspired me to write this exposition.

References

[CT2006] Cover, M. and Thomas, Joy A., Elements of Information Theory 2nd Edition Wiley-Interscience; (2006)

[S1948] Shannon, C. E. A mathematical theory of communication. Bell System Tech. J. 27, (1948). 379423, 623656.

[Y1991] Yeung, Raymond W. A new outlook on Shannon's information measures. IEEE Trans. Inform. Theory 37 (1991), no. 3, part 1, 466474. 94A17

[Y2008] Yeung, Raymond W. Information Theory and Network Coding, Springer (2008)

Department of Mathematics, University of California at San Diego, La Jolla, CA 92093-0112, USA

E-mail address: lrothschild@ucsd.edu